

A Word Shape Coding Method for Camera-based Document Images

[Extended Abstract]

Linlin Li, Chew Lim Tan
School of Computing, National University of Singapore
{lilinlin,tancl}@comp.nus.edu.sg

ABSTRACT

This paper reports a word shape coding method to facilitate retrieval of camera-based document images without OCR. Due to perspective distortion, many reported word shape coding methods fail on camera-based images. In this paper, the problem is addressed by approximating the perspective transformation with an affine transformation, and employing an affine invariant, namely length ratio, to represent the connected components. Components in a document image are classified into a few clusters, each of which is assigned with a representative symbol. Retrieval are based on “words” comprising of symbols. The experiment results showed that the proposed method achieved an average retrieval precision of 93.43% and recall of 94.22%.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*content analysis and indexing*

General Terms

Algorithms

Keywords

Perspective distortion, Document image retrieval

1. INTRODUCTION

Word shape coding is a technique which has been reported for years as a complementary technique to Optical Character Recognition (OCR). Word shape coding is to map a component to a reduced symbol set, other than to give it the exact character identity as OCR does. Thus compared to OCR, word shape coding is much faster in speed and more reliable when image quality is degraded. It has been widely employed in document image processing tasks, such as language identification and duplication detection, on which further indexing and retrieval depend. Besides, it also plays an essential role in document image retrieval when full-scale OCR is too slow or too unreliable to be carried out[1][4][5].

With the rapid advances in camera technology, camera has become an emerging alternative to scanners. Images taken by a camera often suffer from perspective distortion, which never appears in an image produced by a scanner.

Hence, word shape coding techniques developed specifically for scanner-based images no longer work on camera-based images. Also, no reliable OCR techniques for camera-based images are available currently. Thus, a word shape coding method is presented in this paper to efficiently retrieve camera-based document images.

2. APPROACH

The basic assumption of our method is that: for each character in the image, the perspective transformation can be approximated by an affine transformation. Theoretically, when the perceived object depth is much smaller than the distance between the camera lens and the object, the perspective transformation can be approximated by an affine transformation, which can be further decomposed into a scale, a rotation, a shearing, and a translation transformation. Practically, English characters printed on an A4 sheet are generally within a $2 \times 2 \text{ mm}^2$ bounding box. In order to take a photo of the whole sheet, the distance between the camera and the projection center on the sheet is at least 30 mm. In addition, since the image is taken for reading purpose, the camera projection angle is nearly perpendicular to the sheet plane. The object depth of a character is hence very small. Therefore, the affine approximation holds.

2.1 Ratio Histogram

Assume there is a connected component C , such as character ‘A’ in figure 1(a). Firstly, the centroid of the convex image of C , denoted by o , is located. The skeleton of C , as shown in figure 1(b), is gotten by a thinning operation. Then a line through o , shown as the bar in figure 1(c), is arbitrarily drawn. The line will have several intersections with the skeleton. For each pair of intersections, denoted by i_1 and i_2 , the length ratio is calculated as:

$$\begin{cases} \frac{oi_1}{oi_2}, & i_1 \text{ and } i_2 \text{ are at different sides of } o \\ -\frac{oi_1}{oi_2}, & i_1 \text{ and } i_2 \text{ are at the same side of } o \end{cases} \quad (1)$$

where oi_k is the Euclidean distance between o and i_k . In the next step, the line is rotated 360 degrees around o degree by degree, and length ratios are calculated meanwhile. Finally, for each component a histogram is constructed to record the occurrence of length ratios, if their absolute values are larger than 1. In the experiment, a histogram with 20 bins, starting with -10 and ending with 10 was used. In particular, bin 1 kept a record of the number of ratios which are greater than -10 and smaller than -9, and so on and so forth.

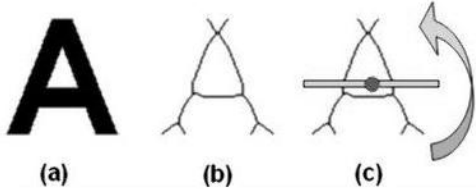


Figure 1: The word shape coding method.

It has been proved that the centroid of a convex polygon preserves under affine transformations[2]. Also, it is well known that the ratio of length of line segments on a given line remains constant under affine transformations. Therefore, the normalized ratio histogram of C will remain unchanged when C is under any affine transformation.

2.2 Document Image Representing

In OCR process, components are assigned with identities in a character set, by being compared with character templates. In our method, components are assigned with identities in a reduced symbol set. The symbol set is found as follows: connected components are extracted from training document images of a fronto-parallel view, and their ratio histograms are calculated. KNN clustering algorithm is applied to histograms with cosine distance as the distance measure. k clusters are found and the centroid of each cluster is recorded as a symbol template. Different characters, which have similar histograms after normalization, are classified into the same cluster, such as ‘o’, ‘c’, and ‘u’. k was chosen as 23 in our experiment, to optimize the separation of clusters.

When an unknown document image is presented, each connected component in the image is compared with the k templates, and its identity is assigned as the identity of the most similar template. However, in the experiment, components smaller than 50 pixels were thrown away in order to avoid pixel quantization errors.

Word boundaries are found by segmentation algorithm [3]. Then “words” comprising of symbols are formed. Thereafter, traditional vector space model with tf.idf representation is applied to those “words”. Similarity between two document images is formulated as:

$$sim(D_u, D_v) = \frac{\sum D_{u,i} D_{v,i}}{\sqrt{\sum D_{u,i}^2 \sum D_{v,i}^2}} \quad (2)$$

where D_u and D_v are documents, and i is the dimension index of the “word” space.

3. EXPERIMENT RESULTS

In order to test the efficiency of the proposed method, the retrieval experiment was designed as follows: 100 pages were selected from 6 documents of the U.S. patent database. These documents were in different categories assigned by the database. We assumed that pages from the same document were similar to each other in content, and different otherwise. These pages were printed out, and 2304×3072 pixel photos were taken by camera. Since it is natural that a printed paper has some warping distortion, this distortion

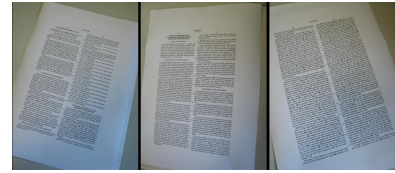


Figure 2: Samples of testing photos.

Doc.	1	2	3	4	5	6
1	0.425	0.079	0.062	0.106	0.117	0.050
2	0.079	0.453	0.108	0.192	0.023	0.151
3	0.062	0.108	0.528	0.088	0.117	0.175
4	0.106	0.192	0.088	0.378	0.066	0.188
5	0.117	0.023	0.117	0.066	0.422	0.138
6	0.050	0.151	0.175	0.188	0.138	0.511

Table 1: Similarity of pages of the same and different documents.

was kept in these photos. It was guaranteed that all characters in a photo were recognizable to people. Examples of the testing data are shown in figure 2. 10 photos were selected as queries. Each query was compared to other 90 photos by the proposed method. When the similarity was greater than a threshold θ , two photos were considered as similar to each other.

In the experiment, an average precision of 93.43% and an average recall of 94.22% were achieved, with $\theta = 0.3$. Table 1 shows the similarity across pages of the same document and that of different documents. Scores were calculated by averaging the similarity between each pair of pages from document i and j . Particularly, the cells on the diagonal are the similarity of pages within the same document. These items are much greater than those off-diagonal items.

Exhaustive study of this method will be done in future. The high recall and precision in our experiment indicate that the technique is also promising for more applications such as language identification and duplication detection.

ACKNOWLEDGMENT: This research is supported by IDM R&D Grant R252-000-325-279.

4. REFERENCES

- [1] D. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3):287–298, 1998.
- [2] C. Gope and N. Kehtarnavaz. Affine invariant comparison of point-sets using convex hulls and hausdorff distances. *Pattern Recognition*, 40(1):309–320, 2007.
- [3] L. Gorman. The document spectrum for page layout analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15:1162–1173, 1993.
- [4] Y. Lu and C. L. Tan. Information retrieval in document image databases. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1398–1410, 2004.
- [5] C. L. Tan, W. Huang, Z. Yu, and Y. Xu. Image document text retrieval without OCR. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24(6):838–844, 2002.